Analysis for the gender discrimination in UC-Berkeley graduate admissions

Zhuoyu Jiang

12/14/2022

Question1

```
#CONSTRUCT THE DATAFRAME
deparment = c(rep('Business Administration',512+313+89+19),
              rep('Physics','History',353+207+17+8),
              rep('History',120+205+202+391),
              rep('English',138+279+131+244),
              rep('Psychology',53+138+94+299),
              rep('Philosophy',22+351+24+317)
)
gender = c(rep('Male',512+313),rep('Female',89+19),
           rep('Male',353+207),rep('Female',17+8),
           rep('Male',120+205),rep('Female',202+391),
           rep('Male',138+279),rep('Female',131+244),
           rep('Male',53+138),rep('Female',94+299),
           rep('Male',22+351),rep('Female',24+317)
)
Admitted = c(rep(1,512),rep(0,313),rep(1,89),rep(0,19),
             rep(1,353),rep(0,207),rep(1,17),rep(0,8),
             rep(1,120),rep(0,205),rep(1,202),rep(0,391),
             rep(1,138),rep(0,279),rep(1,131),rep(0,244),
             rep(1,53),rep(0,138),rep(1,94),rep(0,299),
             rep(1,22),rep(0,351),rep(1,24),rep(0,317)
)
x = data.frame(Admitted,deparment,gender)
head(x)
```

```
##   Admitted               deparment gender
## 1        1 Business Administration   Male
## 2        1 Business Administration   Male
## 3        1 Business Administration   Male
## 4        1 Business Administration   Male
## 5        1 Business Administration   Male
## 6        1 Business Administration   Male
```

```
fit0 = glm(Admitted ~ deparment + gender, data = x, family = binomial)
summary(fit0)
```

```
##
## Call:
## glm(formula = Admitted ~ deparment + gender, family = binomial,
##     data = x)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4773  -0.9306  -0.3741   0.9588   2.3613
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.68192    0.09911   6.880 5.97e-12 ***
## deparmentEnglish    -1.29461    0.10582 -12.234  < 2e-16 ***
## deparmentHistory    -1.26260    0.10663 -11.841  < 2e-16 ***
## deparmentPhilosophy -3.30648    0.16998 -19.452  < 2e-16 ***
## deparmentPhysics    -0.04340    0.10984  -0.395    0.693
## deparmentPsychology -1.73931    0.12611 -13.792  < 2e-16 ***
## genderMale          -0.09987    0.08085  -1.235    0.217
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6044.3  on 4525  degrees of freedom
## Residual deviance: 5187.5  on 4519  degrees of freedom
## AIC: 5201.5
##
## Number of Fisher Scoring iterations: 5
```

The equation is logit(π) = 0.68192 -1.29461 * `deparmentEnglish` -1.26260 * `deparmentHistory` -3.30648 * `deparmentPhilosophy` -0.04340* `deparmentPhysics` -1.73931 * `deparmentPsychology` -0.09987 * `genderMale`, where indicator variables are equal to 1 when their corresponding condition is true and 0 otherwise.

Question2

a.

```
x1 <- x[-1]
x1$yes_prob<- predict(fit0,type = "response")
x1 <- unique(x1)

library(tidyverse)
q2_a <- x %>%
  group_by(deparment, gender, Admitted) %>%
  summarise(
    total = n()
  ) %>%
  pivot_wider(id_cols = c(deparment, gender),
              names_from = Admitted ,
              values_from = total) %>%
  rename(Yes = `1`,
         No = `0`) %>%
  mutate(total = Yes + No) %>%
  left_join(x1) %>%
  mutate(expect_count_Yes = yes_prob*total,
         expect_count_No = (1-yes_prob)*total,
         residual_yes = Yes - expect_count_Yes,
         residual_no = No - expect_count_No)
q2_a
```

```
## # A tibble: 12 x 10
## # Groups:   deparment, gender [12]
##    deparment    gender    No   Yes total yes_p~1 expec~2 expec~3 resid~4 resid~5
##    <chr>        <chr>  <int> <int> <int>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Business Ad~ Female    19    89   108  0.664     71.7   36.3   17.3   -17.3
##  2 Business Ad~ Male     313   512   825  0.642    529.   296.   -17.3    17.3
##  3 English      Female   244   131   375  0.351    132.   243.    -0.793   0.793
##  4 English      Male     279   138   417  0.329    137.   280.     0.793  -0.793
##  5 History      Female   391   202   593  0.359    213.   380.   -10.8    10.8
##  6 History      Male     205   120   325  0.336    109.   216.    10.8   -10.8
##  7 Philosophy   Female   317    24   341  0.0676    23.0  318.     0.957  -0.957
##  8 Philosophy   Male     351    22   373  0.0615    23.0  350.    -0.957   0.957
##  9 Physics      Female     8    17    25  0.654     16.4    8.64   0.640  -0.640
## 10 Physics      Male     207   353   560  0.631    354.   206.    -0.640   0.640
## 11 Psychology   Female   299    94   393  0.258    101.   292.    -7.32    7.32
## 12 Psychology   Male     138    53   191  0.239     45.7  145.     7.32   -7.32
## # ... with abbreviated variable names 1: yes_prob, 2: expect_count_Yes,
## #   3: expect_count_No, 4: residual_yes, 5: residual_no
```

b.

There are two insignificant factor `genderMale` and `departmentPhysics`. To fit the two categories well, the residual of the predict values for other factors in the model enlarges.

```
#The cells for Department Business Administration seem fit bad. The difference between the number of ma
```

c.

```
HLTest = function(obj, g) {
 # first, check to see if we fed in the right kind of object
 stopifnot(family(obj)$family == "binomial" && family(obj)$link == "logit")
 y = obj$model[[1]]
 trials = rep(1, times = nrow(obj$model))
 if(any(colnames(obj$model) == "(weights)"))
  trials <- obj$model[[ncol(obj$model)]]
 # the double bracket (above) gets the index of items within an object
 if (is.factor(y))
  y = as.numeric(y) == 2  # Converts 1-2 factor levels to logical 0/1 values
 yhat = obj$fitted.values
 # browser()
 interval = cut(yhat, unique(quantile(yhat, 0:g/g)), include.lowest = TRUE)  # Creates factor with leve
 Y1 <- trials*y
 Y0 <- trials - Y1
 Y1hat <- trials*yhat
 Y0hat <- trials - Y1hat
 obs = xtabs(formula = cbind(Y0, Y1) ~ interval)
 expect = xtabs(formula = cbind(Y0hat, Y1hat) ~ interval)
 if (any(expect < 5))
  warning("Some expected counts are less than 5. Use smaller number of groups")
 pear <- (obs - expect)/sqrt(expect)
 chisq = sum(pear^2)
 P = 1 - pchisq(chisq, g - 2)
 # by returning an object of class "htest", the function will perform like the
 # built-in hypothesis tests
 return(structure(list(
  method = c(paste("Hosmer and Lemeshow goodness-of-fit test with", g, "bins", sep = " ")),
  data.name = deparse(substitute(obj)),
  statistic = c(X2 = chisq),
  parameter = c(df = g-2),
  p.value = P,
  pear.resid = pear,
  expect = expect,
  observed = obs
 ), class = 'htest'))
}
```

```
HLTest(fit0,g = 10)
```

```
##
##  Hosmer and Lemeshow goodness-of-fit test with 10 bins
##
## data:  fit0
## X2 = 14.815, df = 8, p-value = 0.06284
```

The test statistic is $\chi^2_{HL} = 14.815$ and p-value is 0.06284, which is larger than 0.05. Thus, with 95% confidence interval, we can conclude that the model does not fit bad and we cannot reject the null hypothesis that there is no difference between the expected counts fitted by the model fitted in Problem and each cell in the contingency table.

Question3

a.

```
fit3 = glm(Admitted ~ deparment * gender, data = x, family = binomial)
summary(fit3)
```

```
##
## Call:
## glm(formula = Admitted ~ deparment * gender, family = binomial,
##     data = x)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8642  -0.9127  -0.3821   0.9768   2.3793
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      1.5442     0.2527   6.110 9.94e-10 ***
## deparmentEnglish                -2.1662     0.2750  -7.878 3.32e-15 ***
## deparmentHistory                -2.2046     0.2672  -8.252  < 2e-16 ***
## deparmentPhilosophy             -4.1250     0.3297 -12.512  < 2e-16 ***
## deparmentPhysics                -0.7904     0.4977  -1.588  0.11224
## deparmentPsychology             -2.7013     0.2790  -9.682  < 2e-16 ***
## genderMale                      -1.0521     0.2627  -4.005 6.21e-05 ***
## deparmentEnglish:genderMale      0.9701     0.3026   3.206  0.00135 **
## deparmentHistory:genderMale      1.1770     0.2996   3.929 8.53e-05 ***
## deparmentPhilosophy:genderMale   0.8632     0.4027   2.144  0.03206 *
## deparmentPhysics:genderMale      0.8321     0.5104   1.630  0.10306
## deparmentPsychology:genderMale   1.2523     0.3303   3.791  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6044.3  on 4525  degrees of freedom
## Residual deviance: 5167.3  on 4514  degrees of freedom
## AIC: 5191.3
##
## Number of Fisher Scoring iterations: 5
```

The equation is logit(π) = 1.5442 -2.1662 * `deparmentEnglish` -2.2046 * `deparmentHistory` -4.1250 * `deparmentPhilosophy` -0.04340* `deparmentPhysics` -2.7013 * `deparmentPsychology` -1.0521 * `genderMale` + 0.9701 * `deparmentEnglish:genderMale` +1.1770 * `deparmentHistory:genderMale` + 0.8632 * `deparmentPhilosophy:genderMale` +0.8321 * `deparmentPhysics:genderMale` + 1.2523 * `deparmentPsychology:genderMale`, where indicator variables are equal to 1 when their corresponding condition is true and 0 otherwise.

```
library(lmtest)
lrtest(fit0, fit3)
```

```
## Likelihood ratio test
##
## Model 1: Admitted ~ deparment + gender
## Model 2: Admitted ~ deparment * gender
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -2593.7
## 2  12 -2583.6  5 20.204   0.001144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is 0.001144, which is smaller than 0.05. Thus, we can reject the null hypothesis, and conclude that the model with the interaction situation between department and gender fits the data better.

b.

```
exp(coef(fit3)[7])
```

```
## genderMale
##   0.349212
```

```
exp(coef(fit3)[8:12] + coef(fit3)[7])
```

```
##    deparmentEnglish:genderMale    deparmentHistory:genderMale
##                      0.9212838                      1.1330596
## deparmentPhilosophy:genderMale    deparmentPhysics:genderMale
##                      0.8278727                      0.8025007
## deparmentPsychology:genderMale
##                      1.2216312
```

The AG conditional odds ratios in department Business Administration is 0.349212, which the estimated admission for male is about 0.349212 times than for female in department Business Administration.

The AG conditional odds ratios in department English is 0.9212838 , which the estimated admission for male is about 0.9212838 times than for female in department English.

The AG conditional odds ratios in department History is 1.1330596, which the estimated admission for male is about 1.1330596 times than for female in department History.

The AG conditional odds ratios in department Philosophy is 0.8278727, which the estimated admission for male is about 0.8278727 times than for female in department Philosophy.

The AG conditional odds ratios in department Physics is 0.8025007, which the estimated admission for male is about 0.8025007 times than for female in department Physics.

The AG conditional odds ratios in department Psychology is 1.2216312, which the estimated admission for male is about 1.2216312 times than for female in department Psychology.

c.

The confident level is 0.95.

```
library(mcprofile)
K <-
  matrix(
    c(rep(0,6),1,rep(0,5),
      rep(0,6),1,1,rep(0,4),
      rep(0,6),1,0,1,rep(0,3),
      rep(0,6),1,0,0,1,rep(0,2),
      rep(0,6),1,0,0,0,1,rep(0,1),
      rep(0,6),1,0,0,0,0,1),
    nrow = 6 ,
    byrow = TRUE
  )
K
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
## [1,]    0    0    0    0    0    0    1    0    0     0     0     0
## [2,]    0    0    0    0    0    0    1    1    0     0     0     0
## [3,]    0    0    0    0    0    0    1    0    1     0     0     0
## [4,]    0    0    0    0    0    0    1    0    0     1     0     0
## [5,]    0    0    0    0    0    0    1    0    0     0     1     0
## [6,]    0    0    0    0    0    0    1    0    0     0     0     1
```

```
res_confint <- mcprofile(fit3,
          CM = K)
res_confint
```

```
##
##    Multiple Contrast Profiles
##
##    Estimate Std.err
## C1   -1.052   0.263
## C2   -0.082   0.150
## C3    0.125   0.144
## C4   -0.189   0.305
## C5   -0.220   0.438
## C6    0.200   0.200
```

```
exp(confint(res_confint))
```

```
##
##    mcprofile - Confidence Intervals
##
## level:        0.95
## adjustment:   single-step
##
##    Estimate lower upper
## C1    0.349 0.166 0.671
## C2    0.921 0.620 1.369
## C3    1.133 0.774 1.653
## C4    0.828 0.366 1.857
## C5    0.803 0.228 2.425
## C6    1.222 0.716 2.060
```

The CI of AG conditional odds ratios in department Business Administration is 0.1661733 and 0.6711405, which the estimated admission for male is between 0.1661733 and 0.6711405 times than for female in department Business Administration. Since 1 is not within the confidence interval, we can conclude that with 95% confidence interval, the admission rate of male is less than the female in this department.

The CI of AG conditional odds ratios in department English is 0.6203044 and 1.3685539, which the estimated admission for male is between 0.6203044 and 1.3685539 times than for female in department English. Since 1 is within the confidence interval, we cannot conclude that with 95% confidence interval, the admission rate of male is different from the female in this department.

The CI of AG conditional odds ratios in department History is 0.7742924 and 1.6529059 , which the estimated admission for male is between 0.7742924 and 1.6529059 times than for female in department History.Since 1 is within the confidence interval, we cannot conclude that with 95% confidence interval, the admission rate of male is different from the female in this department.

The CI of AG conditional odds ratios in department Philosophy is 0.3660608 and 1.8566740, which the estimated admission for male is about 0.3660608 and 1.8566740 times than for female in department Philosophy.Since 1 is within the confidence interval, we cannot conclude that with 95% confidence interval, the admission rate of male is different from the female in this department.

The CI of AG conditional odds ratios in department Physics is 0.2282466 and 2.4246334, which the estimated admission for male is between 0.2282466 and 2.4246334 times than for female in department Physics.Since 1 is within the confidence interval, we cannot conclude that with 95% confidence interval, the admission rate of male is different from the female in this department.

The CI of AG conditional odds ratios in department Psychology is 0.7160358 2.0598950 , which the estimated admission for male is between 0.7160358 2.0598950 times than for female in department Psychology.Since 1 is within the confidence interval, we cannot conclude that with 95% confidence interval, the admission rate of male is different from the female in this department.

d.

Except department Business Administration, gender does not have a significant impact on departments' admissions decisions. In department Business Administration, the admission ratio of male is lower than the female.

Question4

a.

```
male.rate=512/313
female.rate=89/19
total.rate=(512+313)/(89+19)
male.rate
```

```
## [1] 1.635783
```

```
female.rate
```

```
## [1] 4.684211
```

```
total.rate
```

```
## [1] 7.638889
```

Yes, in the department Business Administration row of the data, the number of admitting for male is only 1.635783 times than male are not admitted, while there is 4.684211 times for female admitted compared with female not admitted. However, The total number of male applicants is 7.638889 times higher than the total number of female applicants.The reverse of the admission rate and the number of applicants also shows in Physics department. The advantage in admissions ratio disappears under the large difference of the number of applicants. When the departments are combined, Simpson's paradox occurs.

b.

The lawsuit alleged that female applicants were unfairly admitted at a lower rate compared to male applicants. However, we only find that gender does not have a significant impact on departments' admissions decisions except the department Business Administration(the ratio of female student admission is larger than the ratio of male students). Berkeley should realize that they have an imbalance in the number of male and female students they admitted. To avoid the lawsuit, the university should either expand the number of women admitted to the department of Business Administration or the department of Physics, and level up the bar of male admitted in these two departments; or narrow the number of female admitted in other four departments, and enlarge the number male admitted in other four departments to achieve gender balance.